Supplementary material for the submission "Principled Parallel Mean-Field Inference for Discrete Markov Random Fields"

November 13, 2015

In this appendix, we provide more details on the theoretical results presented in the paper. We first recapitulate the problem formulation and notations in Section 1. In Section 2, we derive the update rule of the traditional sweep mean-field method. In Section 3, we provide a detailed derivation of our parallel mean-field update rule. Then, in Section 4, we give prove that our it is guaranteed to converge for the fixed step size. Finally, in Section 5, we provide more details on our method with adaptive step size.

1 Problem Formulation

Recall that mean-field inference solves the following optimization problem:

$$\underset{\mathbf{q}\in\mathcal{M}}{\text{minimize}}\ \mathcal{F}(\mathbf{q})\ , \tag{1}$$

where $\mathcal{F}$ is the variational free energy

$$\mathcal{F}(\mathbf{q}) = \underbrace{-\mathbf{E}_{Q(\boldsymbol{X};\mathbf{q})}[\log P(\mathbf{X}\mid\mathbf{I})]}_{\mathcal{E}(\mathbf{q})} + \underbrace{\mathbf{E}_{Q(\boldsymbol{X};\mathbf{q})}[\log Q(\boldsymbol{X};\mathbf{q})]}_{-\mathcal{H}(\mathbf{q})}, \tag{2}$$

and $Q(\mathbf{X};\mathbf{q})$ is the factorized variational distribution with parameters $\mathbf{q}\in\mathcal{M}$, such that $\forall i,l,\ 0\leq q_{i,l}\leq 1$ and $\forall i, \sum_l q_{i,l}=1$. Q is used to approximate the true posterior $P(\mathbf{X}\mid\mathbf{I})$.

2 Sweep Mean-Field Inference

For completeness, let us first provide the derivation of well-known sweep mean-field updates, similarly to that of [1]. These updates involve minimising of the function $\mathcal{F}(\mathbf{q})$ iteratively with respect to $\mathbf{q}_i=\{q_{i,1},\dots,q_{i,L}\}$, the subset of parameters $\mathbf{q}$ corresponding to the variable X_i. The subset of parameters that correspond to all the other variables, which we will denote by $\mathbf{q}^t_{\neg i}$, remains fixed at the current iteration. We therefore have to

$$\begin{aligned}\underset{\mathbf{q}_i}{\text{minimize}}\quad & \mathcal{E}(\mathbf{q}_i,\mathbf{q}^t_{\neg i}) - \mathcal{H}(\mathbf{q}_i,\mathbf{q}^t_{\neg i})\\ \text{subject to}\quad & \sum_l q_{i,l}=1\ .\end{aligned} \tag{3}$$

Let's first expand the first term of Eq. 3. We write

$$\begin{aligned}\mathcal{E}(\mathbf{q}_i, \mathbf{q}_{\neg i}^t) &= -\mathbf{E}_{Q(\boldsymbol{X};\mathbf{q})}[\log P(\mathbf{X}|\mathbf{I})] \\ &= -\mathbf{E}_{Q(\boldsymbol{X};\mathbf{q})}\left[\mathbf{E}_{Q(\boldsymbol{X}|\mathbf{q})}[\log P(\mathbf{X}|\mathbf{I})|X_i]\right] \\ &= -\sum_l q_{i,l}\mathbf{E}_{Q(\boldsymbol{X};\mathbf{q}_{\neg i})}[\log P(\mathbf{X}|\mathbf{I})|X_i = l]\end{aligned} \tag{4}$$

Since $Q(\mathbf{X};\mathbf{q})$ is a product of categorical distributions $Q_i(\mathbf{X}_i;\mathbf{q})$, we can rewrite the second term of Eq. 3 as

$$\begin{aligned}-\mathcal{H}(\mathbf{q}_i, \mathbf{q}_{\neg i}^t) &= \sum_{j,l} q_{j,l}\log q_{j,l} \\ &= \sum_l q_{i,l}\log q_{i,l} + \underbrace{\sum_{j:j\neq i}\sum_l q_{j,l}\log q_{j,l}}_{C_i} ,\end{aligned} \tag{5}$$

where C_i to denotes the constant summand which does not include terms related to X_i.

Let us now define the Lagrangian

$$\begin{aligned}\mathcal{L}(\mathbf{q}_i, \mu_i) &= \mathcal{E}(\mathbf{q}_i, \mathbf{q}_{\neg i}^t) - \mathcal{H}(\mathbf{q}_i, \mathbf{q}_{\neg i}^t) - \mu_i(\sum_l q_{i,l} - 1) \\ &= -\sum_l q_{i,l}\mathbf{E}_{Q(\boldsymbol{X}|\mathbf{q}_{-i})}[\log p(\mathbf{X}|\mathbf{I})|X_i = l] + \sum_l q_{i,l}\log q_{i,l} - \mu_i(\sum_l q_{i,l} - 1) + C_i .\end{aligned} \tag{6}$$

where we introduced a dual variable μ_i to account for the optimization constraint. By differentiating with respect to a $q_{i,l}$ we obtain the optimality condition

$$\log q_{i,l}^\star = \mathbf{E}_{Q(\boldsymbol{X}|\mathbf{q}_{\neg i})}[\log p(\mathbf{X}|\mathbf{I})|X_i = l] + \mu_i . \tag{7}$$

This leads to the standard update rule

$$\forall l \ , q_{i,l}^\star \propto \exp\left[\mathbf{E}_{Q(\boldsymbol{X}|\mathbf{q}_{\neg i})}[\log p(\mathbf{X}|\mathbf{I})|X_i = l]\right] \ , \tag{8}$$

where the normalization constant can be computed from μ_i.

Iteratively applying 8 then guarantees the convergence of $\mathcal{F}$, due to the fact that $\mathcal{F}$ is convex with respect to each $q_{i,l}$ [1].

3 Proximal Gradient Mean-Field Inference

We will now derive the closed-form update rule for the KL-proximal gradient descent introduced in Section 3.1 of the paper.

Let us now consider the proximal gradient update,

$$\underset{q\in\mathcal{M}}{\text{minimize}}\left\{\langle \mathbf{q}, \nabla\mathcal{E}(\mathbf{q}^t)\rangle - \mathcal{H}(\mathbf{q}) + \mathbf{D}^t \odot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t)\right\} \ , \tag{9}$$

where the first and the second terms are the expected energy and negative entropy respectively, and the last term is the proximal term. It can be written as

$$\mathbf{D}^t \odot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t) = \sum_{i,l} d_{i,l} \cdot q_{i,l} \log \frac{q_{i,l}}{q^t_{i,l}} \ , \tag{10}$$

where $\mathbf{D}^t$ is a diagonal matrix with non-zero elements $d_{i,l}$.

Our goal is to derive a closed-form update for all the mean parameters $q_{i,l}$, or, alternatively, for all the natural parameters $\theta_{i,l}$. By using Eq. 4, we can write down the partial derivative of the expected energy with respect to any $q_{i,l}$ as

$$\nabla\mathcal{E}(\mathbf{q}^t)_{i,l} = \frac{\partial\mathcal{E}(\mathbf{q}^t)}{\partial q_{i,l}} = \mathbf{E}_{Q(\mathbf{X}|\mathbf{q}^t_{\neg i})}[\log p(\mathbf{X}|\mathbf{I})|X_i = l] \ . \tag{11}$$

Note, that both our objective $\mathcal{F}$ *and* the constraints $\mathbf{q} \in \mathcal{M}$ are separable over the variables $X_1, \ldots, X_N$, which makes it possible to minimize independently for each X_i. In other words, our goal is to solve for all i

$$\underset{\mathbf{q}_i}{\text{minimize}} \qquad \sum_l q_{i,l}\nabla\mathcal{E}(\mathbf{q}^t)_{i,l} + \sum_l q_{i,l}\log q_{i,l} + d^t_i \sum_l q_{i,l} \log \frac{q_{i,l}}{q^t_{i,l}} \ , \tag{12}$$

$$\text{subject to} \qquad \sum_l q_{i,l} = 1 \tag{13}$$

Similarly to the sweep updates decribed in Section 2, we convert each problem to an unconstrained one by introducing the Lagrangian

$$\begin{aligned}\mathcal{L}(\mathbf{q}_i, \mu_i) = &\sum_l q_{i,l}\nabla\mathcal{E}(\mathbf{q}^t)_{i,l} + \sum_l q_{i,l}\log q_{i,l} \ , \\ &+ d^t_i \sum_l q_{i,l} \log \frac{q_{i,l}}{q^t_{i,l}} - \mu_i \left(\sum_l q_{i,l} - 1\right) \ ,\end{aligned} \tag{14}$$

where μ_i is a corresponding Lagrange multiplier.

We then differentiate it with respect to $q_{i,l}, \ \forall i, l$

$$(1 + d^t_i)\log q^\star_{i,l} = \mathbf{E}_{Q(\boldsymbol{X}|\mathbf{q}_{-i})}[\log p(\mathbf{X}|\mathbf{I})|X_i = l] + d^t_i \log q^t_{i,l} + \mu_i \ , \tag{15}$$

which in turn leads to the update rule

$$q^{t+1}_{i,l} \propto \exp\left[\eta^t_i \cdot \mathbf{E}_{Q(\boldsymbol{X}|\mathbf{q}_{-i})}[\log p(\mathbf{X}|\mathbf{I})|X_i = l] + (1 - \eta^t_i) \cdot \log q^t_{i,l}\right] \ , \tag{16}$$

where $\eta^t_i = \frac{1}{1+d^t_i}$, and normalization constant can be obtained from μ_i.

4 Proving Convergence

We will now prove that our fixed step-size algorithm guarantess convergence. In the remainder of the supplementary material, we will work under the assumption that

$$\forall i, t \ \ \exists d^t_i \text{ s.t } \forall l \ d^t_{i,l} = d^t_i,$$

which is verified for the fixed and adaptive step size and methods described in the paper. We will therefore replace $d_{i,l}$ by d_i in the subsequent derivations. Note that, this property does not hold for `OURS-ADAM`. Nevertherless, as shown in the experimental evaluation, in practice it tends to converge faster and to a better minima.

Lemma 4.1 *The gradient of the proximal term at the current iteration point $\nabla_{\mathbf{q}}\mathbf{D}^t \odot KL(\mathbf{q}||\mathbf{q}^t)|_{\mathbf{q}=\mathbf{q}^t}$ is orthogonal to $\mathcal{M}$.*

Proof Let's write down the gradient:

$$\nabla_{\mathbf{q}}\mathbf{D}^t \odot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t) = (d_1^t \cdot \nabla_{\mathbf{q}_1}\mathrm{KL}(\mathbf{q}_1||\mathbf{q}_1^t), \ldots, d_N^t \nabla_{\mathbf{q}_N}\mathrm{KL}(\mathbf{q}_N||\mathbf{q}_N^t)) \ , \tag{17}$$

with each component containing:

$$\nabla_{\mathbf{q}_i}\mathrm{KL}(\mathbf{q}_i||\mathbf{q}_i^t) = (\log\frac{q_{i,1}}{q_{i,1}^t} + 1, \ldots, \log\frac{q_{i,M}}{q_{i,M}^t} + 1) \ . \tag{18}$$

The partial gradient at the current iteration point $\mathbf{q}_i^t$ is the all-ones vector:

$$\nabla_{\mathbf{q}_i}\mathrm{KL}(\mathbf{q}_i||\mathbf{q}_i^t)|_{\mathbf{q}_i=\mathbf{q}_i^t} = (1, \ldots, 1) \ , \tag{19}$$

which is obviously orthogonal to the hyperplane defined by the constraint $\sum_l q_{i,l} = 1$. Thus, $d_i^t \nabla_{\mathbf{q}_i}\mathrm{KL}(\mathbf{q}_i||\mathbf{q}_i^t)|_{\mathbf{q}_i=\mathbf{q}_i^t}$ is also orthogonal to this hyperplane, and we easily obtain the orthogonality of the product vector $\nabla_{\mathbf{q}}\mathbf{D}^t \odot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t)|_{\mathbf{q}=\mathbf{q}^t}$ to $\mathcal{M}$.

Lemma 4.2 *For all $\mathbf{q}^t$ in $\mathcal{M}$,*

$$\forall \mathbf{q} \in \mathcal{M}, \ \mathbf{D}^t \cdot KL(\mathbf{q}^{t+1}||\mathbf{q}^t) \geq \frac{L}{2}\|\mathbf{q} - \mathbf{q}^t\|_2^2 \ .$$

Proof Note that the Hessian of the KL-proximal term is diagonal with

$$\forall \mathbf{q} \in \mathcal{M}, \ \frac{\partial^2 \mathbf{D}^t \cdot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t)}{\partial q_{i,l}^2}|_{\mathbf{q}} = \frac{d_{i,l}^t}{q_{i,l}} \geq L \ . \tag{20}$$

Therefore, the proximal term is L-strongly convex on $\mathcal{M}$. For all $\mathbf{q}^t$ in $\mathcal{M}$,

$$\forall \mathbf{q} \in \mathcal{M}, \ \mathbf{D}^t \cdot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t) \geq \langle \nabla_{\mathbf{q}}\mathbf{D}^t \odot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t)|_{\mathbf{q}=\mathbf{q}^t}, \mathbf{q} - \mathbf{q}^t \rangle + \frac{L}{2}\|\mathbf{q} - \mathbf{q}^t\|_2^2 \ . \tag{21}$$

The first term of the right hand side is null according to the orthogonality property 4.1. Which leads to

$$\forall \mathbf{q} \in \mathcal{M}, \ \ \mathbf{D}^t \cdot \mathrm{KL}(\mathbf{q}^{t+1}||\mathbf{q}^t) \geq \frac{L}{2}\|\mathbf{q} - \mathbf{q}^t\|_2^2 \ . \tag{22}$$

We will now demonstrate, that under certain assumptions, applying updates of Eq. 16 lead to a decrease in objective at each iteration.

Theorem 4.3 *If $\mathcal{E}$ is L-Lipschitz gradient on $\mathcal{M}$, and that d_i^ts are chosen such that $d_i^t \geq L, \ \forall t, i$. Then the objective function is decreasing at each step.*

Proof Let us assume that $\mathcal{E}$ is L-Lipschitz gradient on $\mathcal{M}$ and that $d_i^t \geq L, \ \forall t, i$. Then, we can show that the value of the objective function $\mathcal{E}(\mathbf{q}^{t+1}) - \mathcal{H}(\mathbf{q}^{t+1})$ at step $t+1$ has to be smaller than $\mathcal{E}(\mathbf{q}^t) - \mathcal{H}(\mathbf{q}^t)$

$$\mathcal{E}(\mathbf{q}^t) - \mathcal{H}(\mathbf{q}^t) \geq \operatorname*{argmin}_{\mathbf{q}} \left[\mathcal{E}(\mathbf{q}^t) + \langle(\mathbf{q} - \mathbf{q}^t), \nabla\mathcal{E}(\mathbf{q}^t)\rangle - \mathcal{H}(\mathbf{q}) + \mathbf{D}^t \cdot \mathrm{KL}(\mathbf{q}||\mathbf{q}^t)\right] \tag{23}$$

$$\geq \mathcal{E}(\mathbf{q}^t) + \langle(\mathbf{q}^{t+1} - \mathbf{q}^t), \nabla\mathcal{E}(\mathbf{q}^t)\rangle - \mathcal{H}(\mathbf{q}^{t+1}) + \mathbf{D}^t \cdot \mathrm{KL}(\mathbf{q}^{t+1}||\mathbf{q}^t) \tag{24}$$

$$\geq \mathcal{E}(\mathbf{q}^t) + \langle(\mathbf{q}^{t+1} - \mathbf{q}^t), \nabla\mathcal{E}(\mathbf{q}^t)\rangle - \mathcal{H}(\mathbf{q}^{t+1}) + \frac{L}{2}\|\mathbf{q}^{t+1} - \mathbf{q}^t\|_2^2 \tag{25}$$

$$\geq \mathcal{E}(\mathbf{q}^{t+1}) - \mathcal{H}(\mathbf{q}^{t+1}) \tag{26}$$

where step Eq. 24 comes from the fact that by definition $\mathbf{q}^{t+1}$ realizes the minimum, Eq. 25 holds by strong-convexity lower bound 4.2 and Eq. 26 holds by L-Lipschitz gradient property of $\mathcal{E}$.

5 Adaptive Steps

We now formally justify the update rule used in Section 3.3 of the paper. In the proof of Lemma 4.2, in Eq. 20, we used the fact that $\frac{1}{q_{i,l}} \geq 1$. This bound is correct, but, often very large since that $q_{i,l}$ can be very close to 0. This leads to the choice $d_i = L$, for all i, which ensures $\frac{d_i}{q_{i,l}} \geq L$.

An alternative, is to choose a smaller value $d_i = L \max(q^t_{i,0}, \ldots, q^t_{i,L_i-1})$, which also ensures that $\frac{d_i}{q^t_{i,l}} \geq L$ for all i, l, but the gain is very marginal.

However, all the previous bounds ignore the fact that all our variables lie on the simplex $\mathcal{M}$. We will now show, that we can obtain a proximal term that locally upper-bounds the objective function much more closely.

We start by writing a second order Taylor expansion of the KL-proximal term for variable i around the current iteration point. This yields

$$d_i^t \mathrm{KL}(\mathbf{q}^{t+1}||\mathbf{q}^t) = d_i^t \langle \nabla_{\mathbf{q}_i} \mathrm{KL}(\mathbf{q}_i||\mathbf{q}_i^t)|_{\mathbf{q}_i=\mathbf{q}_i^t}, \mathbf{q}_i^{t+1} - \mathbf{q}_i^t\rangle + \frac{d_i^t}{2} \sum_l \frac{(q_{i,l}^{t+1} - q_{i,l}^t)^2}{q_{i,l}^t} + o(\|\mathbf{q}_i^{t+1} - \mathbf{q}_i^t\|_2^2) \tag{27}$$

$$= \frac{d_i^t}{2} \sum_l \frac{(q_{i,l}^{t+1} - q_{i,l}^t)^2}{q_{i,l}^t} + o(\|\mathbf{q}_i^{t+1} - \mathbf{q}_i^t\|_2^2) \; . \tag{28}$$

where we applied Lemma 4.1 to get Eq. 28.

For a derivation similar to Eq. 23-Eq. 26 to hold (up to a second order approximation), we just need to choose d_i^t so that $d_i^t \sum_l \frac{(q_{i,l}^{t+1} - q_{i,l}^t)^2}{q_{i,l}^t} \geq L\|\mathbf{q}_i^{t+1} - \mathbf{q}_i^t\|_2^2$.

However, we should take into account the fact that $\mathbf{q}^{t+1}$ and q^t lie in $\mathcal{M}$, and therefore $\sum_l q_{i,l}^{t+1} - q_{i,l}^t = 0$. Therefore, one can choose $\frac{L}{d_i^t}$ as the optimum of the following program:

$$\begin{aligned} \underset{\delta}{\text{minimize}} \quad & \sum_l \frac{\delta_l^2}{q_{i,l}^t} \; , \\ \text{subject to} \quad & \sum_l \delta_l = 0 \; , \\ & \sum_l \delta_l^2 = 1 \; . \end{aligned} \tag{29}$$

Finding an efficient way to obtain solutions to this program for general label size is left for future work. For binary variables, it is easy to show that the optimum of the program above is $\frac{1}{2q_{i,0}^t q_{i,1}^t}$.

This is why, we choose $d_i^t = dq_{i,0}^t q_{i,1}^t$ in Section 3.3 of the paper.

References

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 1, 2